Central Limit Theorem Compared to Exponential Distribution in R -- Part 1

Douglas Wirtz

February 24, 2016

Overview

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

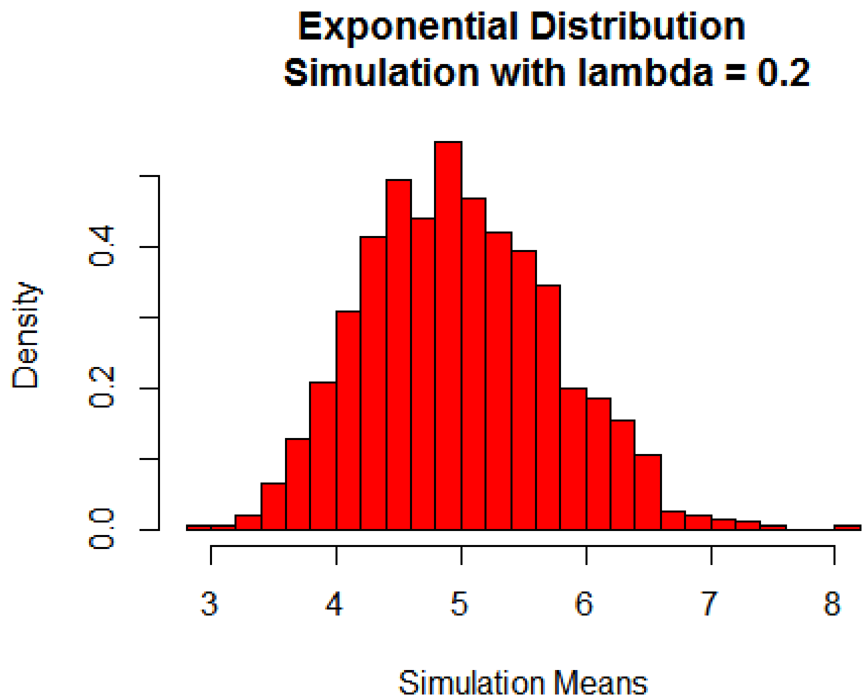
Simulations

This code shows the simulations needed for data analysis.

```r
#Set variables from instructions and seed for reproducibility
set.seed(18)
lambda <- 0.2
n <- 40
sim_num <- 1000
#Run simulation
sim <- matrix(rexp(sim_num * n, lambda), sim_num, n)
row_means <- rowMeans(sim)
```

Plot the distribution of the simulation means

```r
hist(row_means, breaks = 25, col = "red", probability = TRUE,
     main = "Exponential Distribution
     Simulation with lambda = 0.2",
     xlab = "Simulation Means")
```



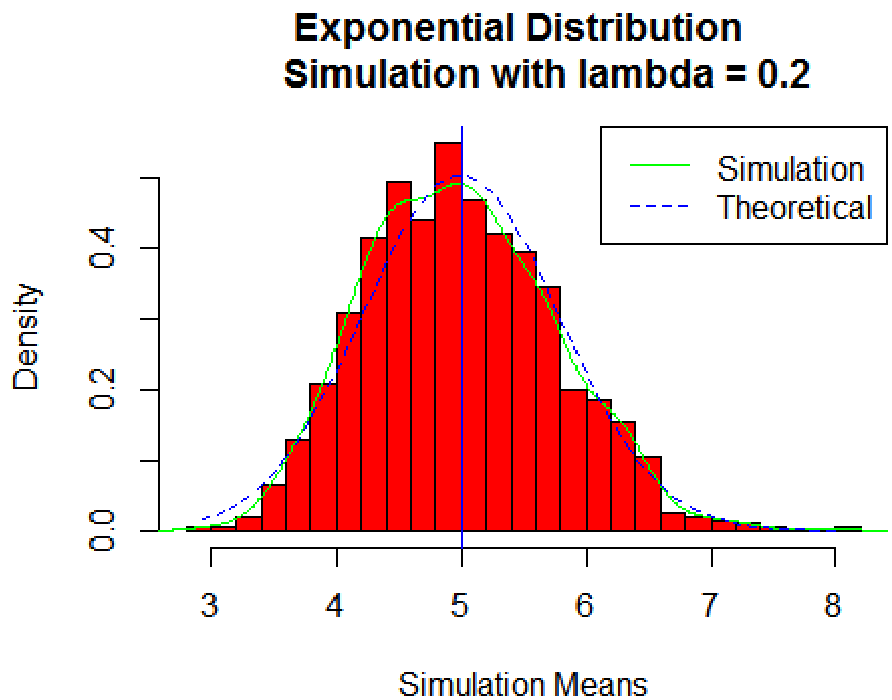
This plot shows the distribution of the sample means after 1000 simulations.

Sample Mean vs Theoretical Mean

The sample mean and theoretical mean are calculated and then plotted on the distribution to show comparison

```
sim_mean <- mean(row_means)
sim_mean

## [1] 4.9928

theoretical_mean <- 1/lambda
theoretical_mean

## [1] 5

hist(row_means, breaks = 25, col = "red", probability = TRUE,
     main = "Exponential Distribution
     Simulation with lambda = 0.2",
     xlab = "Simulation Means")
lines(density(row_means), col = "green")
abline(v = theoretical_mean, col = "blue")
xfit <- seq(min(row_means), max(row_means), length = 100)
yfit <- dnorm(xfit, theoretical_mean, theoretical_mean/sqrt(n))
lines(xfit, yfit, col = "blue", lty = 2)
```

```
legend('topright', c("Simulation", "Theoretical"),
       lty = c(1,2), col = c("green", "blue"))
```



The sample mean of 4.9928 compared to the theoretical mean of 5 implies that the simulation distrubution and the theoretical normal distriution are very close.

Sample Variance vs Theoretical Variance

The sample and theoretical variance are calculated to show comparison

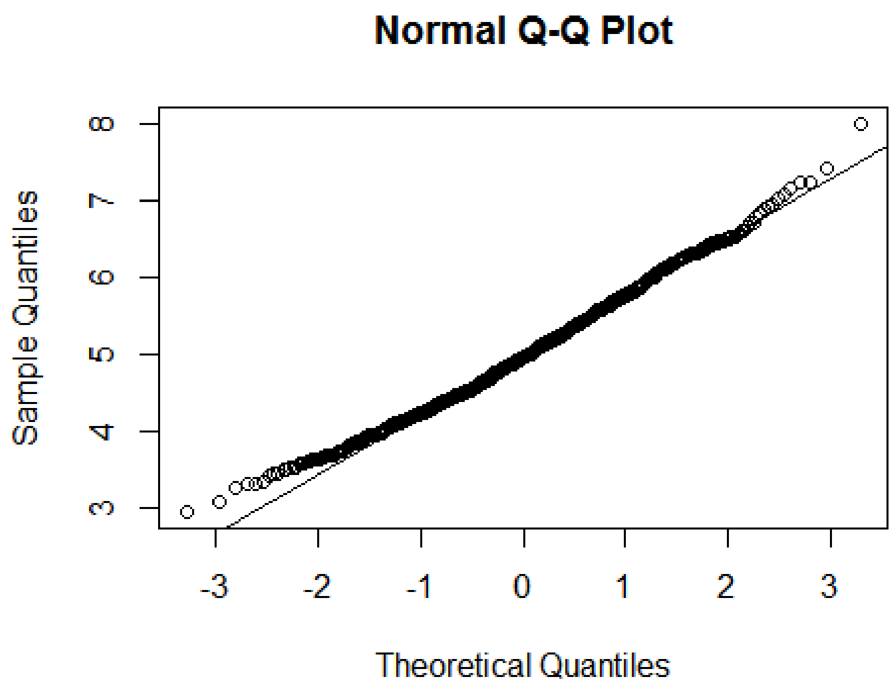
```
sim_var <- var(row_means)
sim_var

## [1] 0.5718996

theoretical_var <- (1/lambda)^2/n
theoretical_var

## [1] 0.625
```

After calculating the sample variance of 0.5718996 and the theoretical variance of 0.625, again, they seem to be very close.

Is the Distribution Appoximately Normal?

Due to the central limit theorem, the averages of samples follow normal distribution. The figure above shows a normally distributed simulation with the simulation and theoretical means plotted on top to show normality. A comparison of the simulation and theoretical variances also suggest a normal distribution. In addition, a q-q plot, shown below, shows the theoretical quantiles matching closely with the simulation quantiles.

```
qqnorm(row_means)
qqline(row_means)
```



As another piece of evidence to show normality, the confidence intervals for the simulation and theoretical means are calculated below.

```
sim_CI <- sim_mean + c(-1,1)*1.96*sd(row_means)/sqrt(n)
sim_CI

## [1] 4.758439 5.227161

theoretical_CI <- theoretical_mean + c(-1,1)*1.96
*sqrt(theoretical_var)/sqrt(n)
theoretical_CI

## [1] 4.755 5.245
```

As you can see, the results from the confidence interval calculation show that the intervals are essentially identical and therefore, the simulation distribution is apporximately normal.